Coursera Capstone Project: Opening a new Gym

Willem Wissing

1. Introduction

1.1 Background

Opening a new gym comes with many challenges, such as obtaining the right equipment, hiring skilled staff, and finding a location or venue. On top of that there is the competition to keep in mind.

1.2 Problem

Finding the right location can be tricky. What makes for a good location for a gym, and where is the competition the least? If we want to start a gym in one of the Netherlands' biggest cities, where should we start looking?This project aims to identify locations or areas where a gym might do well.

2. Data

2.1 Data sources

To get a list of the 20 biggest cities in the Netherlands we scraped this wikipedia page. In order to get a list of defined areas we decided to use the Geonames API to find all the postcodes in each city. Finally, we used the Foursquare API to locate all gyms and other venues in each area.

2.2 Data cleaning

The data scraped and requested from APIs were mostly in good shape, but some cleaning was still required.

First the list of cities contained some other data related to images on the wikipedia page, which had to be removed. Further, one of the city names (the Hague) did not match the name used in the Geonames api ('s-Gravenhage, one of the dutch names for this city) and had to be converted.

The dutch postal code system uses a 4 digit number and 2 characters to uniquely identify every street in the country. Considering this would be too specific, we chose to discard the character identifiers and just use the 4 digit numbers that identify each area (this was later done directly in the request to Geonames). Of the remaining results, 2 city names are also the names of the provinces they are located in. Therefore our request for these was filled with areas not just in the city, but the

whole province. By matching on municipality we was able to filter out these unwanted results.

Foursquare was used to located gyms and sporting facilities within 500 meters of each of our locations. This resulted in some duplicates, as the venues were close to 2 different locations. While we kept this data for the analysis of the individual areas, we temporarily removed these duplicates when generating city-wide statistics.

3. Methodology

In order to find the right location we are going to focus on 2 features: Population/gym density per city, and amount of gyms per area compared to similar areas.

People don't necessarily go to the gym that is in their neighborhood, or even the closest to them. We assume however that people do tend to stay within the same city, and the relative amount of gym-goers per capita is similar for each city. Therefore a city with fewer gyms relative to their population would make a better target than an already gym-saturated environment.

To find the right location within these cities, we want to how many gyms each area typically has, and whether there is room for more. In order to do this, we use the foursquare data for all venues found per area. We then try to cluster these areas based on their most common venues, and calculate the average amount of sporting facilities for each cluster.

4. Analysis

First we look at the total amount of postal codes found per city. A quick look suggests that the larger cities seem to have the most area codes, which is as we expect. Next we try to plot our area locations on a map of the Netherlands, so we can visually see our distribution.

Place	
Amsterdam	81
Rotterdam	79
's-Gravenhage	61
Utrecht	46
Almere	43
Groningen	43
Eindhoven	34
Haarlemmermeer	33
Apeldoorn	33
's-Hertogenbosch	30
Tilburg	28
Arnhem	26
Zwolle	25
Breda	25
Nijmegen	25
Zaanstad	23
Enschede	23
Haarlem	20
Amersfoort	18
Leiden	16

Figure 1: Amount of area codes per city

Figure 2: A visual of the distribution of area codes

When we look at a zoomed in graph of Amsterdam, we can see a lot of overlap in location areas (the blue circle has a radius of 500 meter, the same as our Foursquare search criteria will be). We don't think this is a problem for the analysis of individual locations, as venues that are in a different area code but close by are still relevant to our analysis.

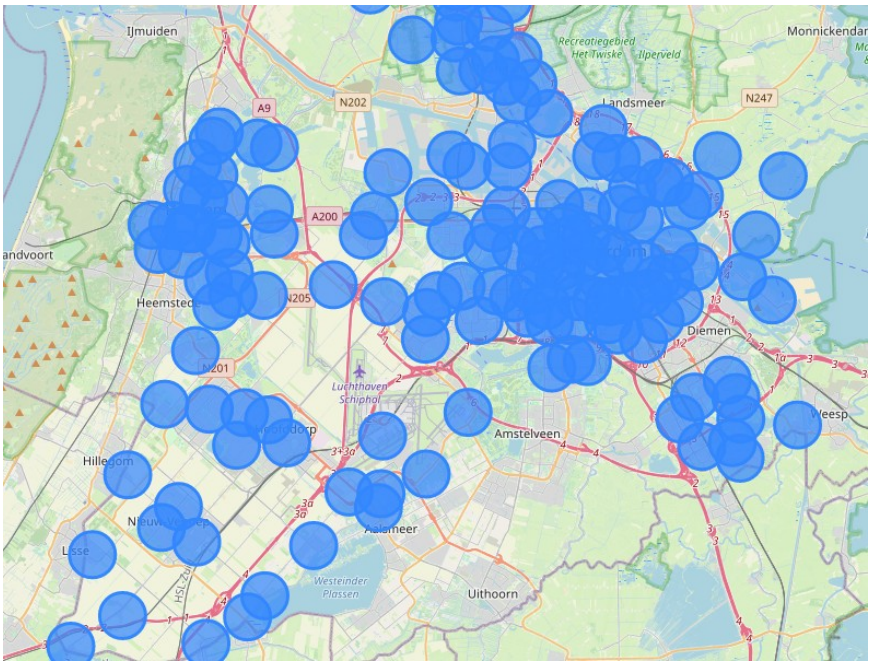

Figure 3: Areas in Amsterdam

When we plot our gym locations in this graph, we get an idea how they are distributed throughout the cities. We can already see a little bit of clustering, which indicates that some areas are more suited than others.

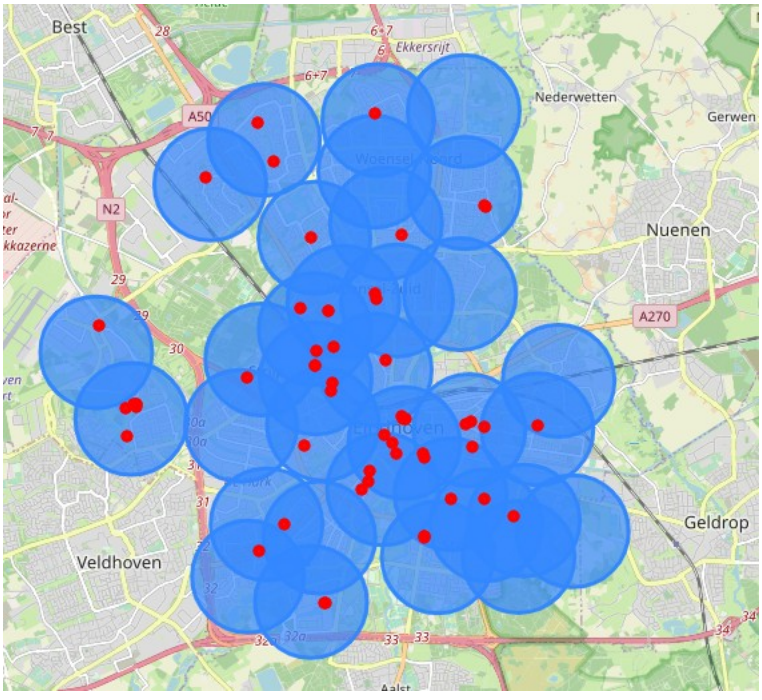

Figure 4: In red, the location of gyms and sporting facilities in Eindhoven

After request all other types of venues, we can check how many we find per city. Again we can see that the biggest cities have the most venues as we expect. Now we will group these venues by postal code and try to find the most common venues.

City	
's-Gravenhage	2773
's-Hertogenbosch	605
Almere	757
Amersfoort	529
Amsterdam	4656
Apeldoorn	553
Arnhem	718
Breda	563
Eindhoven	1000
Enschede	500
Groningen	1246
Haarlem	947
Haarlemmermeer	640
Leiden	616
Nijmegen	818
Rotterdam	3321
Tilburg	729
Utrecht	1913
Zaanstad	422
Zwolle	538

Figure 5: Total amount of venues found per city

	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	1011	Hotel	Bar	Coffee Shop	Restaurant
1	1012	Hotel	Bar	Coffee Shop	Restaurant
2	1013	Bar	Restaurant	Coffee Shop	Café
3	1014	Restaurant	Nightclub	Music Venue	Italian Restaurant
4	1015	Hotel	Bar	Café	Coffee Shop

Figure 6: Most common venues per postal code

Now we have our data we can try to identify similarity between areas. Using the a Kmeans algorithm we divide the location data up into 5 different groups, based on the 10 most common venues in their area.

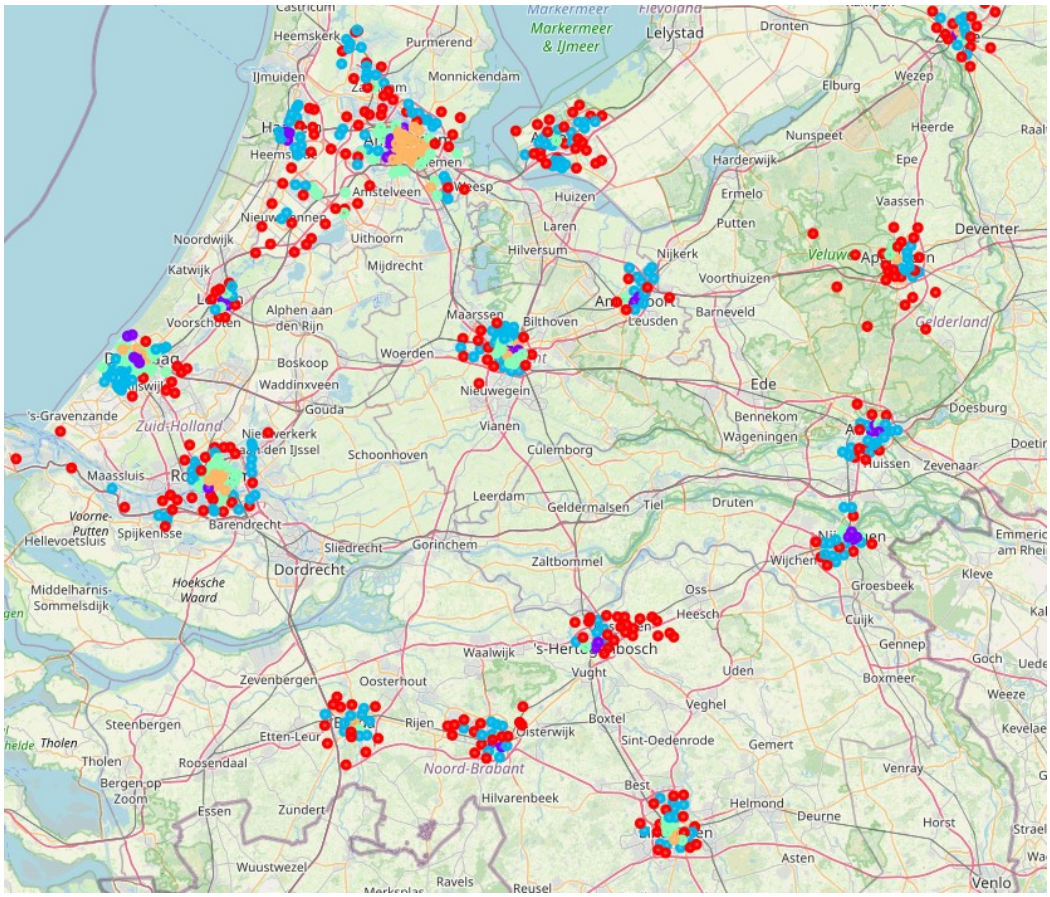

Figure 7: Clustered location data

From this map we can see that areas outside of a city tend to be similar, and diversity is really found in the centre.

Now we plot the location data for our gym venues in this map.

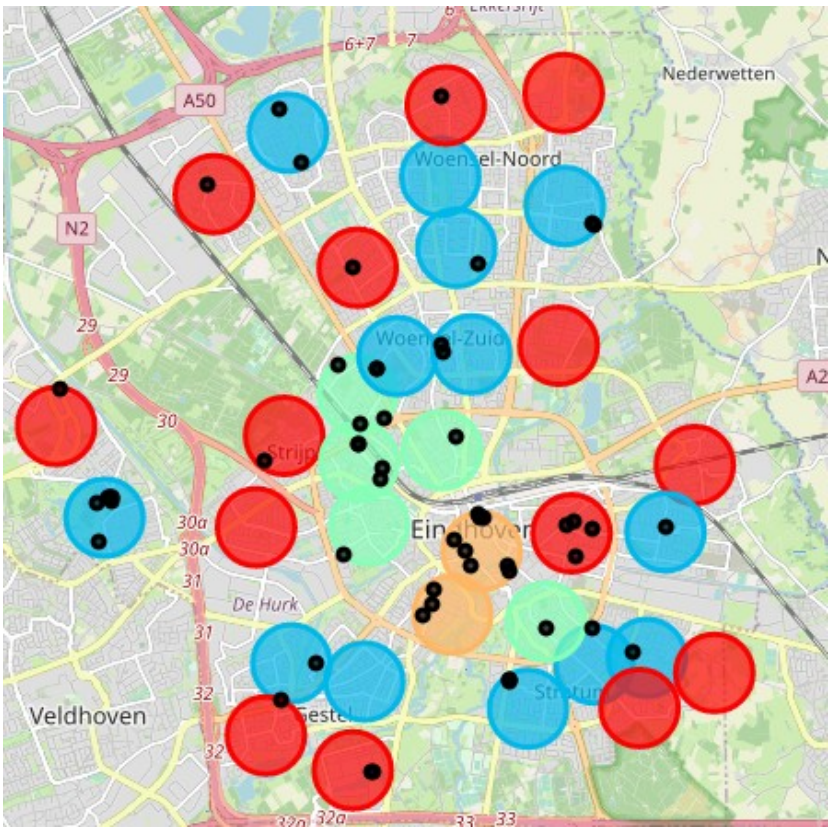

Figure 8: Gym locations (black) and Clustered areas in Eindhoven

5. Results

The following results were found. The top 3 cities with the fewest gyms per capita in the netherlands are: Enschede, ‘s-Hertogenbosch and Haarlemmermeer, with Amersfoort a close 4th. Next we see a list of postal codes in these cities, that have fewer gyms in them than their counterparts based on similarity.

	City	Population	Total Gyms	gym density
1	Haarlem	162,864	65	2505.600000
3	Utrecht	357,179	134	2665.514925
9	Leiden	125,434	44	2850.772727
0	Amsterdam	872,680	294	2968.299320
8	Zwolle	128,617	38	3384.657895
2	's-Gravenhage	544,766	158	3447.886076
5	Zaanstad	156,703	44	3561.431818
5	Groningen	232,826	62	3755.258065
7	Almere	211,514	55	3845.709091
6	Tilburg	219,632	57	3853.192982
0	Apeldoorn	163,706	42	3897.761905
9	Nijmegen	177,818	45	3951.511111
1	Rotterdam	650,711	164	3967.750000
8	Breda	184,403	45	4097.844444
2	Arnhem	161,260	39	4134.871795
4	Eindhoven	234,235	56	4182.767857
4	Amersfoort	157,286	35	4493.885714
6	Haarlemmermeer	155,770	34	4581.470588
7	's-Hertogenbosch	154,989	32	4843.406250
3	Enschede	159,934	31	5159.161290

Figure 9: Population/gym density per city

	Postal Code	Cluster Labels	Venue	Venue Average	Municipality
0	5211	3.0	4.0	4.363636	's-Hertogenbosch
1	5247	2.0	1.0	2.239216	's-Hertogenbosch
2	5215	2.0	2.0	2.239216	's-Hertogenbosch
3	5391	2.0	1.0	2.239216	's-Hertogenbosch
4	5242	0.0	1.0	2.704545	's-Hertogenbosch
5	5243	2.0	1.0	2.239216	's-Hertogenbosch
6	5221	2.0	1.0	2.239216	's-Hertogenbosch
7	5234	2.0	2.0	2.239216	's-Hertogenbosch
8	5241	2.0	1.0	2.239216	's-Hertogenbosch
9	5233	2.0	1.0	2.239216	's-Hertogenbosch
10	5235	0.0	2.0	2.704545	's-Hertogenbosch
11	5213	0.0	1.0	2.704545	's-Hertogenbosch
12	5236	0.0	2.0	2.704545	's-Hertogenbosch
13	5237	2.0	2.0	2.239216	's-Hertogenbosch
14	5223	2.0	2.0	2.239216	's-Hertogenbosch
15	5232	2.0	1.0	2.239216	's-Hertogenbosch
16	7546	0.0	1.0	2.704545	Enschede
17	7548	2.0	1.0	2.239216	Enschede
18	7531	2.0	1.0	2.239216	Enschede
19	7545	0.0	1.0	2.704545	Enschede
20	7514	4.0	4.0	4.016129	Enschede
21	7511	4.0	4.0	4.016129	Enschede
22	7535	2.0	2.0	2.239216	Enschede
23	7513	0.0	2.0	2.704545	Enschede
24	7544	0.0	2.0	2.704545	Enschede
25	7543	0.0	1.0	2.704545	Enschede
26	7512	0.0	2.0	2.704545	Enschede
27	7521	0.0	2.0	2.704545	Enschede
28	1171	0.0	1.0	2.704545	Haarlemmermeer
29	2134	2.0	2.0	2.239216	Haarlemmermeer
30	1119	2.0	1.0	2.239216	Haarlemmermeer
31	2131	2.0	2.0	2.239216	Haarlemmermeer
32	2154	2.0	1.0	2.239216	Haarlemmermeer
33	2144	2.0	1.0	2.239216	Haarlemmermeer

Figure 10: Prime gym locations

6. Discussion

Our final result is a list of 40 different postal codes that could be a good location to open a gym. While this should be a good starting point for further investigation, there are improvements to be made. The postal codes used do not all cover similarly sized areas, making comparison and Foursquare queries more difficult. Also we look mostly at the area in which the gyms are located, not their metric distance to each other. If one were to convert the positional data to metric distances we could generate a more accurate heat map of distance to the closest gym.

Furthermore, a good location for a new venue is much more than just distance to other gyms and location. Real estate prices, available venues and area populations all come in to play. With this data a more in depth research could be done, but is out of scope for this project.

7. Conclusion

We've seen that our top 3 cities are 's-Hertogenbosch, Enschede and Haarlemmermeer, and for each city we've located several areas which are prime candidates. With these starting points one should be able to find a suitable place to start their new enterprise, or expand an existing one.

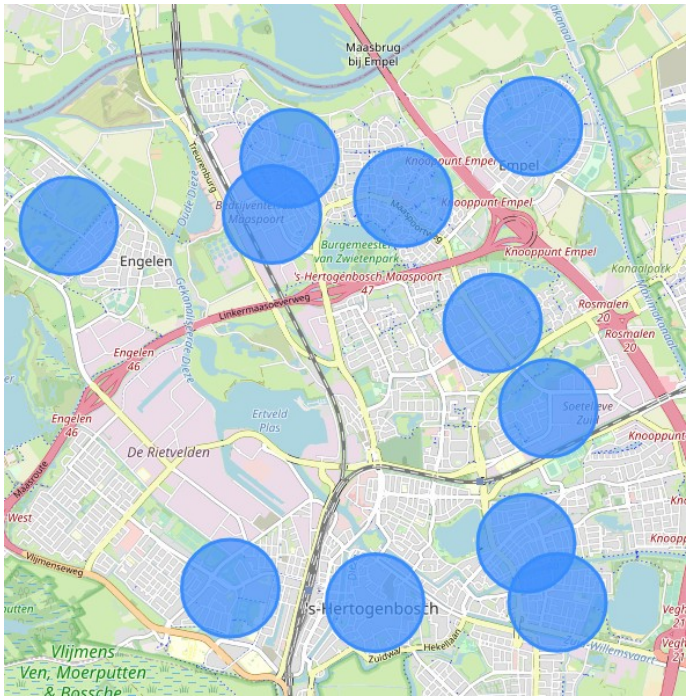

Figure 11: Prime locations in Den Bosch